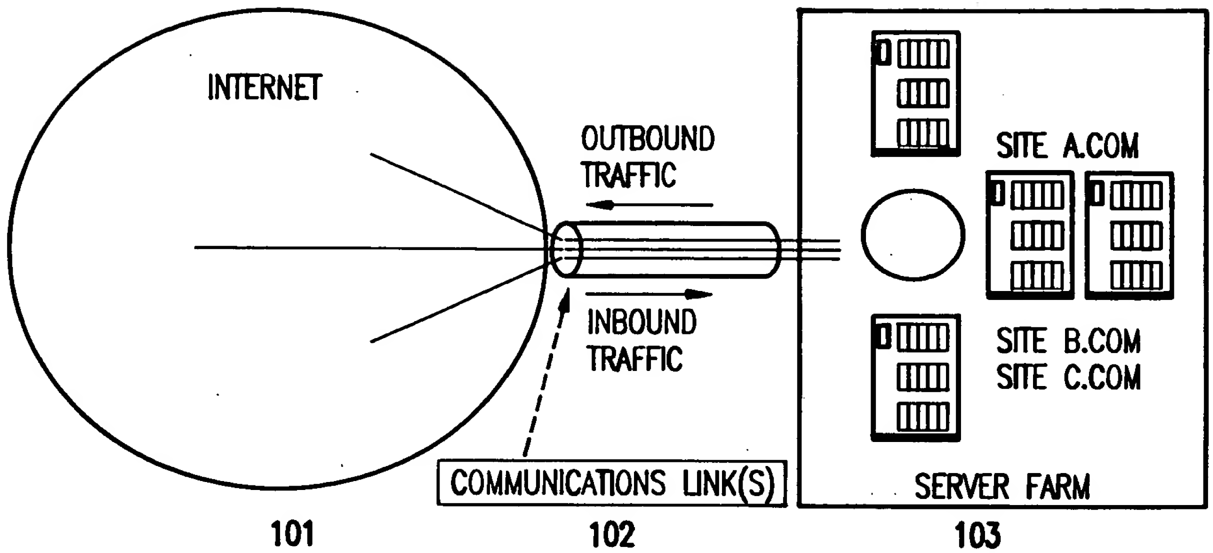
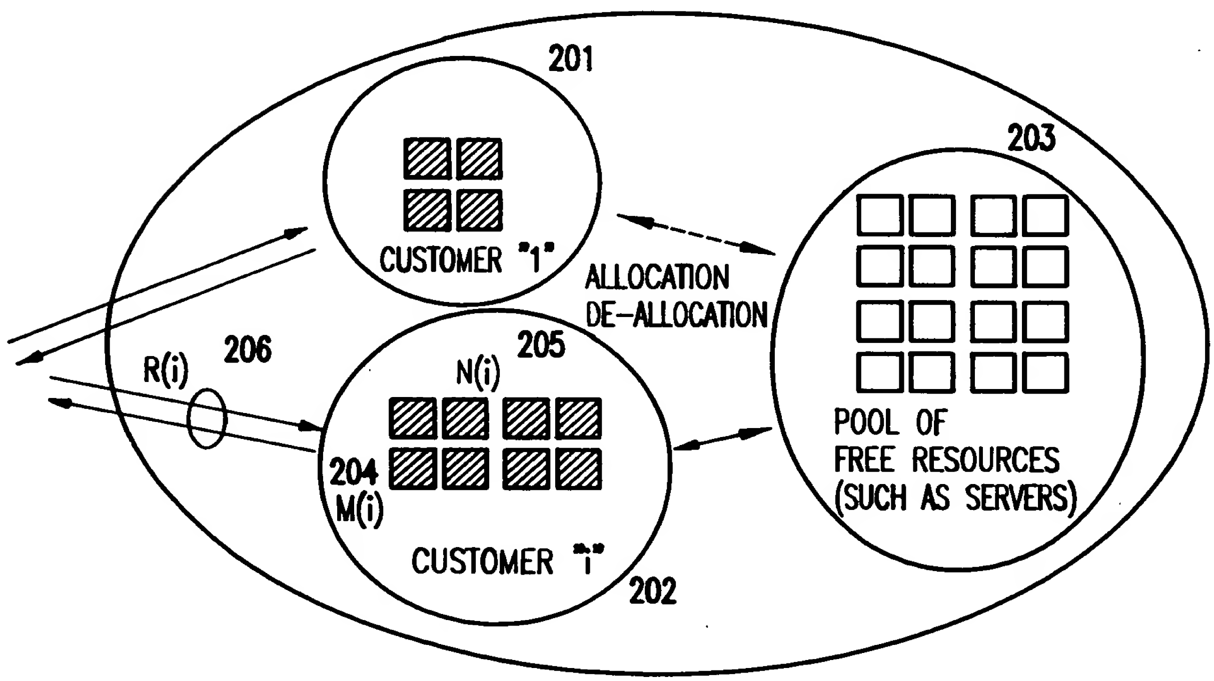


FIG.1



MEASURE R(i), N(i), AND M(i)
COMPUTE Nt(i) AND Rt(i) FROM Mt(i), R(i), N(i) AND M(i),
AND THEN IF NEEDED, MOVE M(i) TO Mt(i) BY EITHER CHANGING N(i) TO Nt(i) OR CHANGING R(i) TO Rt(i).

FIG.2

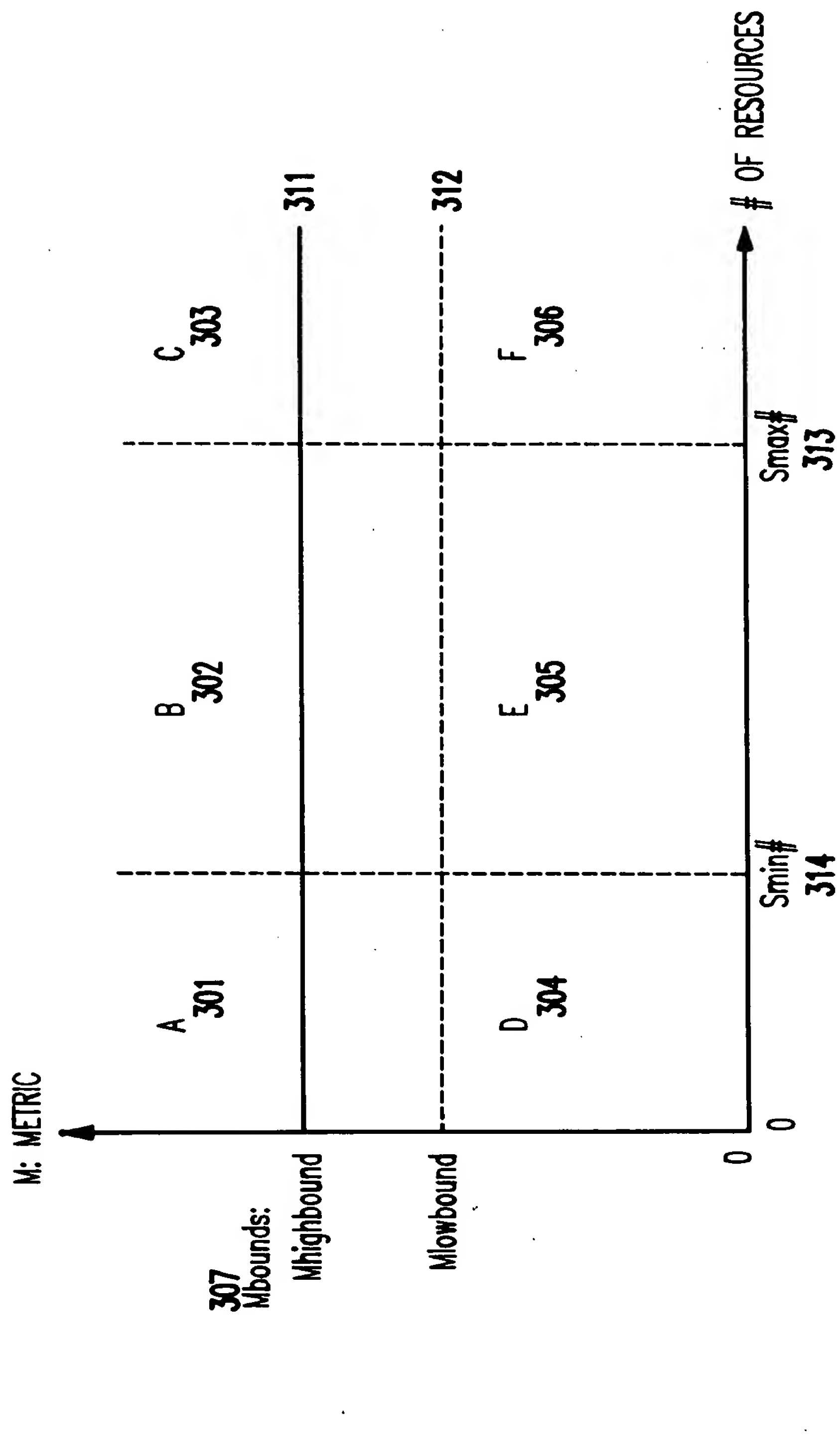

M: METRIC
307
Mbounds:
Mhighbound
Mlowbound
0
0
A
301
B
302
C
303
D
304
E
305
F
306
311
312
Smin#
314
Smax#
313
OF RESOURCES

FIG.3

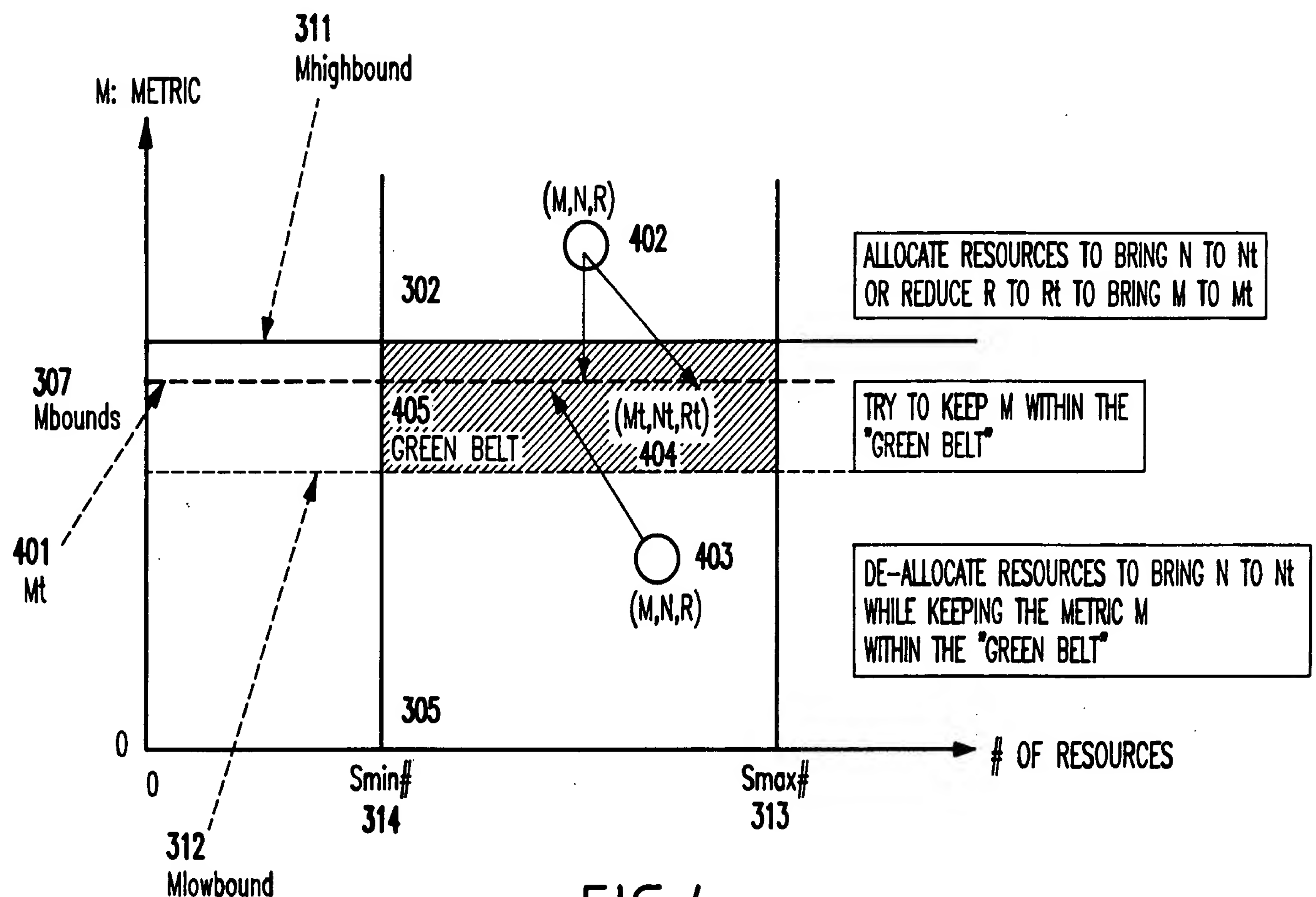

M: METRIC
311
Mhighbound
307
Mbounds
401
Mt
312
Mlowbound
0
0
302
305
(M,N,R)
402
(Mt,Nt,Rt)
404
405
GREEN BELT
403
(M,N,R)
Smin#
314
Smax#
313
OF RESOURCES
ALLOCATE RESOURCES TO BRING N TO Nt
OR REDUCE R TO Rt TO BRING M TO Mt
TRY TO KEEP M WITHIN THE
"GREEN BELT"
DE-ALLOCATE RESOURCES TO BRING N TO Nt
WHILE KEEPING THE METRIC M
WITHIN THE "GREEN BELT"

FIG.4

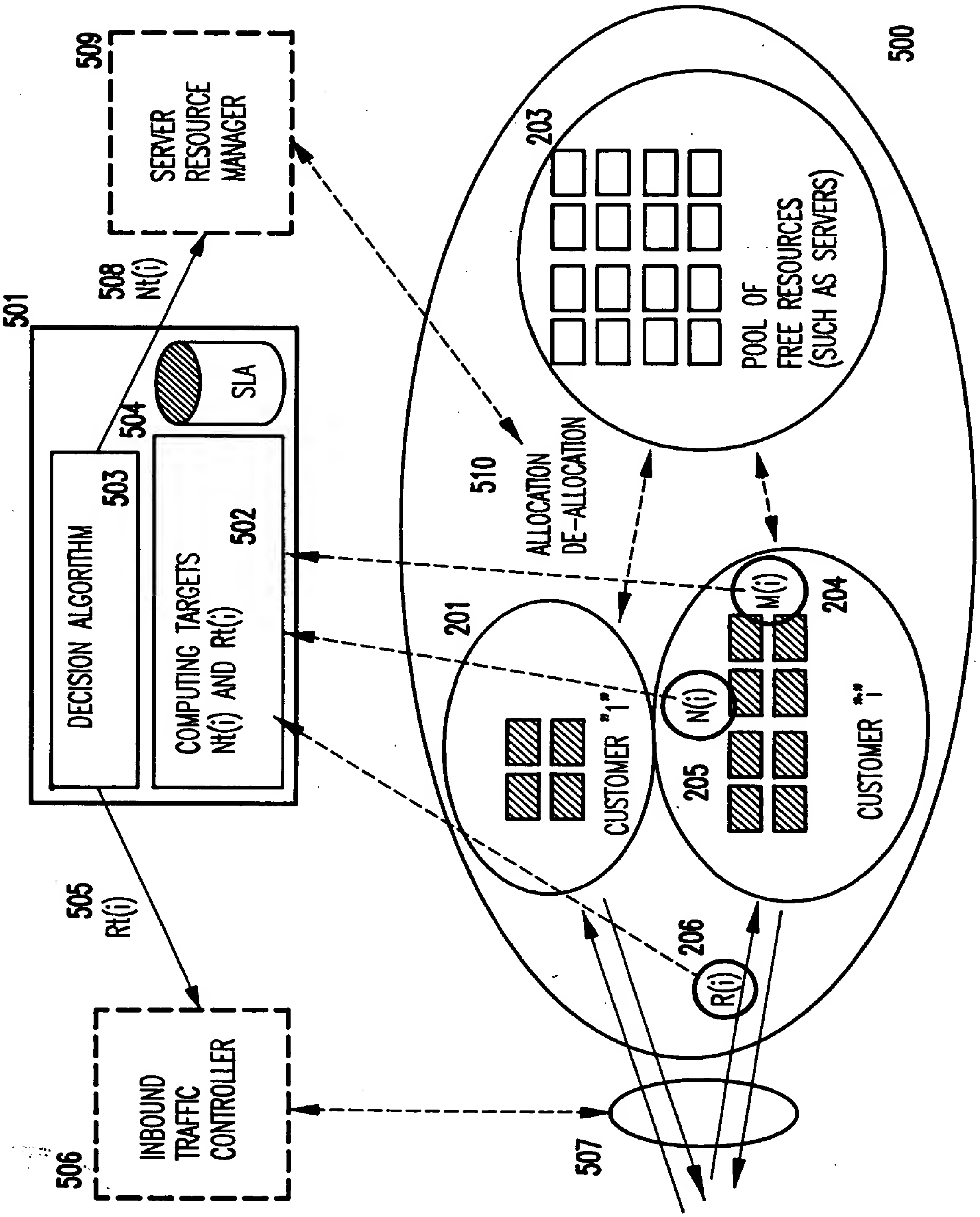

501
DECISION ALGORITHM
503
COMPUTING TARGETS
Nt(i) AND Rt(i)
502
504
SLA
508
Nt(i)
509
SERVER
RESOURCE
MANAGER
505
Rt(i)
506
INBOUND
TRAFFIC
CONTROLLER
507
510
ALLOCATION
DE-ALLOCATION
203
POOL OF
FREE RESOURCES
(SUCH AS SERVERS)
201
CUSTOMER "1"
204
CUSTOMER "i"
M(i)
N(i)
205
206
R(i)
500

FIG.5

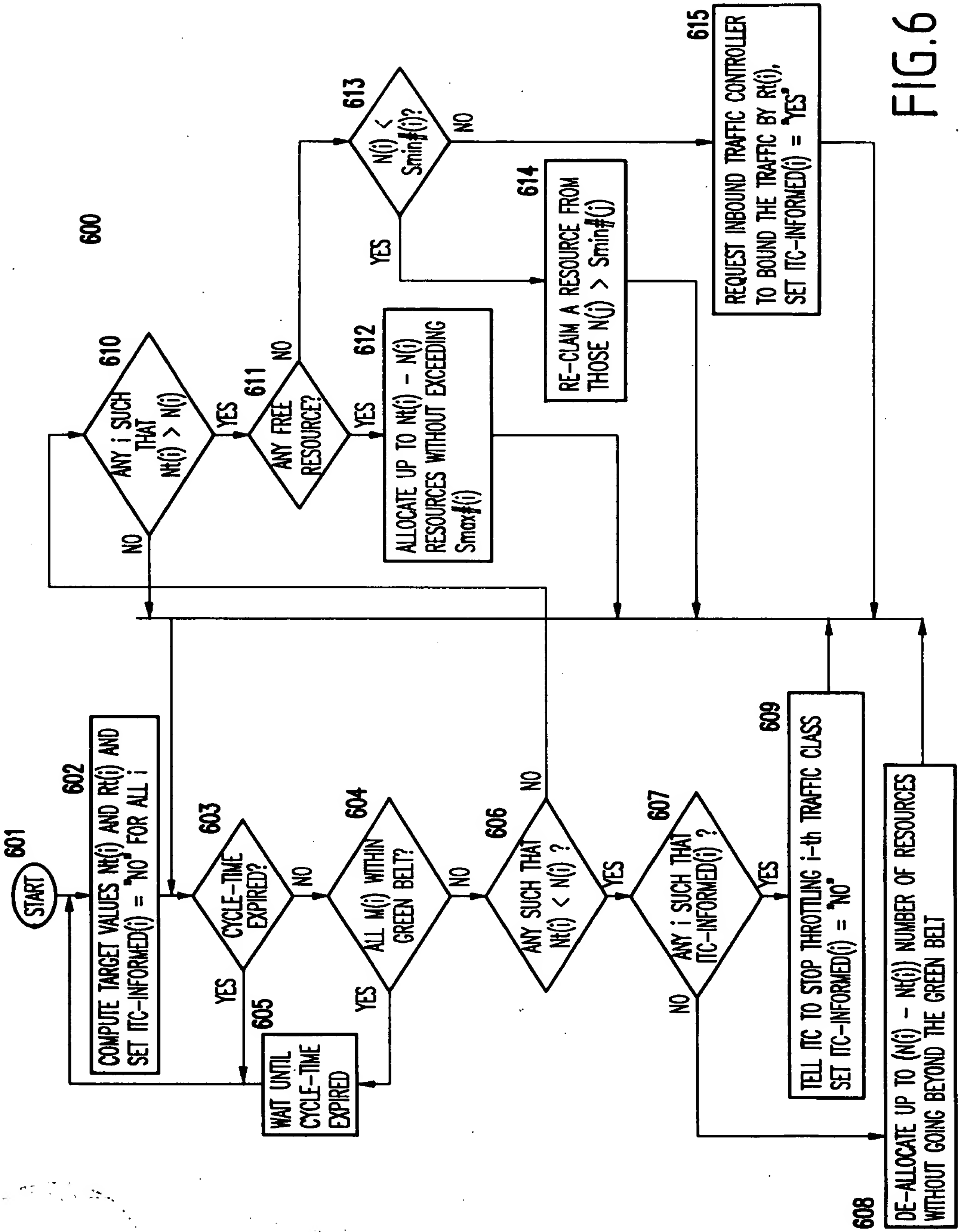

600
601
START
602
COMPUTE TARGET VALUES Nt(i) AND Rt(i) AND SET ITC-INFORMED(i) = "NO" FOR ALL i
603
CYCLE-TIME EXPIRED?
YES
NO
604
ALL M(i) WITHIN GREEN BELT?
YES
NO
605
WAIT UNTIL CYCLE-TIME EXPIRED
606
ANY SUCH THAT Nt(i) < N(i) ?
YES
NO
607
ANY i SUCH THAT ITC-INFORMED(i) ?
YES
NO
608
DE-ALLOCATE UP TO (N(i) - Nt(i)) NUMBER OF RESOURCES WITHOUT GOING BEYOND THE GREEN BELT
609
TELL ITC TO STOP THROTTLING i-th TRAFFIC CLASS SET ITC-INFORMED(i) = "NO"
610
ANY i SUCH THAT Nt(i) > N(i)
YES
NO
611
ANY FREE RESOURCE?
YES
NO
612
ALLOCATE UP TO Nt(i) - N(i) RESOURCES WITHOUT EXCEEDING Smax#(i)
613
N(i) < Smin#(i)?
YES
NO
614
RE-CLAIM A RESOURCE FROM THOSE N(j) > Smin#(j)
615
REQUEST INBOUND TRAFFIC CONTROLLER TO BOUND THE TRAFFIC BY Rt(i), SET ITC-INFORMED(i) = "YES"

FIG.6

Smin#(i) : the amount of resources guaranteed for the i-th customer. This can be a vector.

Smax#(i) : the maximum amount of service resources that could be made available to the i-th customer. This can be a vector.

Mbounds(i) : the bounds on the service level metric.
Each "bounds" consists of a pair, "highbound" and "lowbound."

Ubounds(i) : the bound on the utilization of resources allocated to the i-th customer

Tbounds(i) : the bound on the agreed upon average server response time for the i-th customer

T%bounds(i) : the bound on the agreed upon server response time percentile for the i-th customer

(Smin#(i),Smax#(i),Mbound(i)) : the SLA supported by the invention

N(i) : the number (or amount of) of resources currently allocated to the i-th customer.

R(i) : the current inbound traffic rate for the i-th customer. This could be a vector when more than one type of traffic is defined for each customer.

M(i) : the current value of the metric M for the i-th customer. This could be a vector.
Examples are:
- **U(i)** : the current utilization of the allocated resources to the i-th customer
- **T(i)** : currently observed server response time averaging for the I-th customer
- **T%(i)** : currently observed server response time percentile for the I-th customer

Mt(i) : the "target" (want to achieve) metric value for the i-th customer. Its dimension is same as the dimension of M(i).
This is within the defined "green belt" which is the region within which M(i) is kept.
Examples of Mt(i) are:
- **Ut(i)** : the target resource utilization when M = U,
- **Tt(i)** : the target average response time when M = T
- **Tt%(i)** : the target percentile response time when M = T%

Table 1

For Utilization as Metric: M = U and Mt = Ut

The following relationships hold among various variables:
U(i) = C(i)R(i) / N(i), where C(i) is a constant
Ut(i) = C(i)R(i) / Nt(i), and
Ut(i) = C(i)Rt(i) / N(i).

From the above and from the given values of N(i), R(i), U(i), and the target value Ut(i), Nt(i) and Rt(i) can be computed as follow:

Nt(i) = CEILING [N(i)U(i) / Ut(i)], and
Rt(i) = FLOOR [R(i)Ut(i) / U(i)],

where CEILING gives the smallest integer exceeding and FLOOR gives the largest integer not exceeding.

Table 2

For Average Response Time as Metric: M = T and Mt = Tt

S(i) :server "service" (or processing) time for the i-th customer, this can be computed from observing each individual server service time, or estimated from a queueing formula:

S(i) is a function of {T(i),R(i),N(i)}
If the cluster of servers is modeled by the M/M/m queueing system,
S(i) = ((R(i)T(i)+N(i)+p{N(i)}) - SQRT((R(i)T(i)+N(i)+p{N(i)})**2 - 4R(i)T(i)R(i) / 2R(i)
where p{m} is the probability that there are m requests in the i-th customer's server cluster

For the M/M/m queuing model,
Tt(i) ~ S(i) + p{Nt(i)}S(i) / (Nt(i)-R(i)S(i))
Tt(i) ~ S(i) + p{N(i)}S(i) / (N(i)-Rt(i)S(i))
Therefore,
Nt(i) = CEILING [R(i)S(i) + p{Nt(i)}S(i) / (Tt(i)-S(i))]
Rt(i) = FLOOR [N(i)/S(i) - p{N(i)}/(Tt(i)-S(i))]

where p{m} is the probability that there are m requests in the customer's server cluster

Table 3

For Percentile Response Time as Metric: M=T% and Mt=Tt%

If T%(i) > T%bound(i), then the average response time T(i) needs to be reduced by (T%(i) - T(i)). Therefore, for T%(i) to approach T%bound, the average response time target Tt(i) becomes:

$$Tt(i) = T(i) - (T\%(i) - T\%bound(i)).$$

For the M/M/m queueing model,

$$Tt(i) \sim S(i) + p\{Nt(i)\}S(i) / (Nt(i)-R(i)S(i))$$

$$Tt(i) \sim S(i) + p\{N(i)\}S(i) / (N(i)-Rt(i)S(i))$$

and thus,

$$\mathbf{Nt(i)} = \text{CEILING}\ [R(i)S(i) + p\{Nt(i)\}S(i) / (Tt(i)-S(i))]$$

$$\mathbf{Rt(i)} = \text{FLOOR}\ [N(i)/S(i) - (p\{N(i)\}/Tt(i)-S(i))]$$

where p{m} is the probability that there are m requests in the customer's server cluster

Table 4

For any given metric M,

There are quick simulation tools, quick numerical computation tools and other approximation formula are available in computing **Nt(i)** and **Rt(i)** from given (i.e., measured) values of R(i), N(i) and M(i).

Table 5